Prediction of Used Car Price

Goal:

This project aims to use a machine learning algorithm to predict the price of used cars based on features such as mileage, model year, brand, accident history, etc.

Dataset:

We use the "Used Car Price Prediction Dataset" from Kaggle (https://www.kaggle.com/datasets/taeefnajib/used-car-price-prediction-dataset), which contains 4009 data points. Key features include:

- **Brand & Model:** Identifies the manufacturer and model of the vehicle, categorized into **luxury** (e.g., Tesla, Mercedes-Benz) and **economic** brands (e.g., Toyota, Ford).
- **Model Year:** Indicates the year of manufacture, important for assessing depreciation, technological advancements, and the vehicle's age.
- **Mileage:** Represents the distance traveled by the vehicle, used to estimate wear and tear.
- **Fuel Type:** Specifies whether the car uses gasoline, diesel, electric, or hybrid fuel.
- **Engine:** Includes details on engine power (horsepower), fuel injection type, cylinder count, and engine capacity.
- **Transmission:** Includes the transmission type (manual, automatic) and number of gears (e.g., 6-speed).
- **Exterior & Interior Colors:** Explore the aesthetic aspects of the vehicles, including exterior and interior color options.
- **Accident History:** Indicates whether the car has been involved in any accidents.
- **Clean Title:** Shows whether the vehicle has a clean legal title.
- **Price:** The target variable represents the vehicle's listed price.

Stakeholders:

- **Used Car Dealers:** To determine the best selling price.
- **Car Owners:** To set a fair price for their vehicles.
- **Online Marketplaces:** To facilitate smoother transactions.
- **Insurance Companies:** To calculate premiums based on vehicle value.
- **Banks/Financial Institutions:** To assess loan amounts based on vehicle valuation.

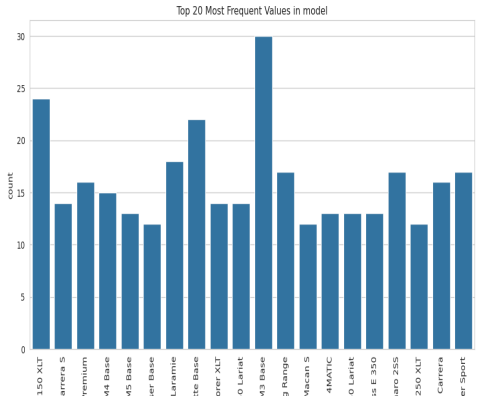
KPIs:

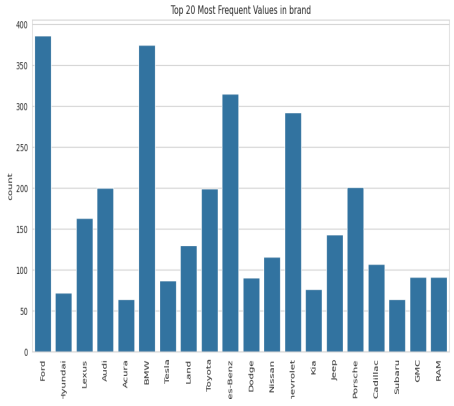
- **Mean Squared Error (MSE):** Measures the average squared difference between predicted and actual prices, with lower values indicating better performance.
- **R-Squared (R^2):** Shows the proportion of variance explained by the model. A higher R^2 indicates a better fit.
- **Mean Absolute Error (MAE):** Measures the average absolute difference between predicted and actual prices for practical, real-world accuracy.

Exploratory Analysis:

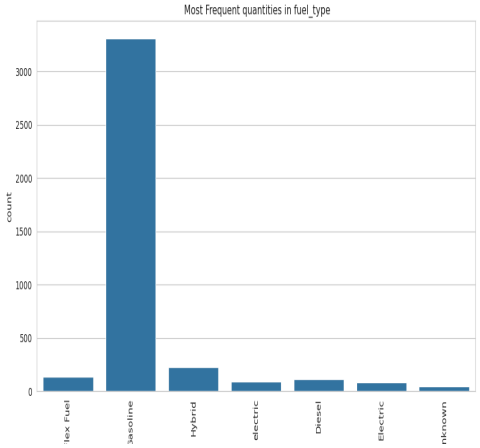
Navdeep:

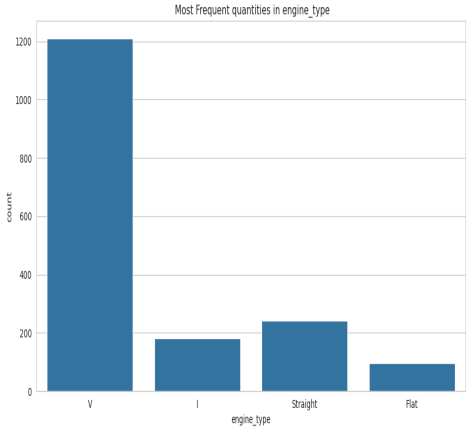
1. **Top brand:** Out of 56 brands, Ford has the most sold used cars in this sample, having a price range starting at $3,000 and going all the way up to $4,20,000.





2. **Frequent fuel type:** The most common fuel type in the dataset is **gasoline** and most frequent engine type is **vertical**.

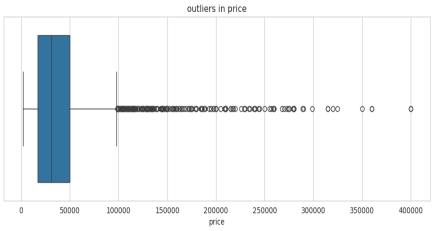


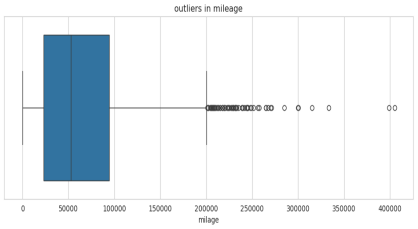


3. **Outliers in the price:** Outliers detected using IQR and various percentiles show a decreasing trend in outlier count as higher percentiles are used. Starting with 244 outliers at the IQR upper limit of $99,175, the number drops to 41 at the 99th percentile, where the price threshold reaches $272,713. However, further exploration is necessary to understand the implications of this cap on model performance and the overall distribution of car prices.

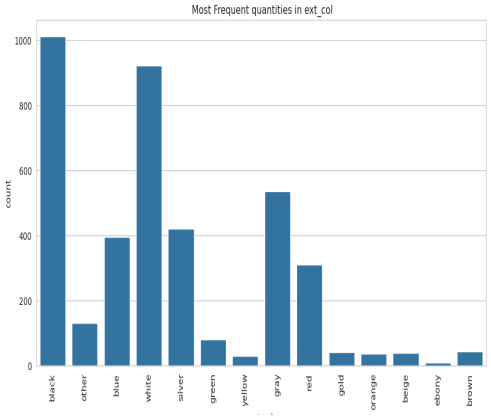
4. **Outliers in the mileage:** Outliers detected using IQR and percentile methods show a decreasing trend as the mileage thresholds increase. Starting with 69 outliers at the IQR upper limit (200,684 miles), the number drops to 41 at the 99th percentile (mileage of 222,428), with no outliers beyond the 100th percentile, where the maximum mileage is 405,000. The observations suggest that the distribution of mileage in the dataset has a long

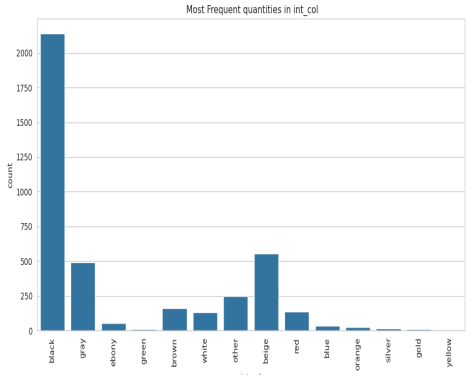
tail of high values, similar to the price distribution in the previous case.

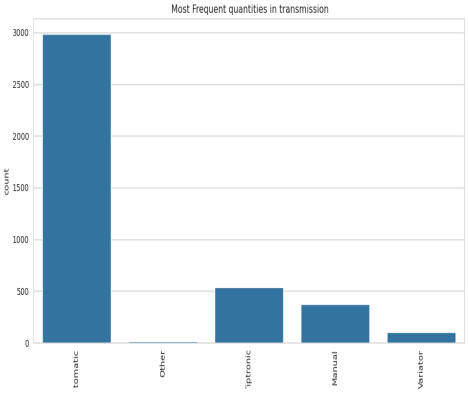


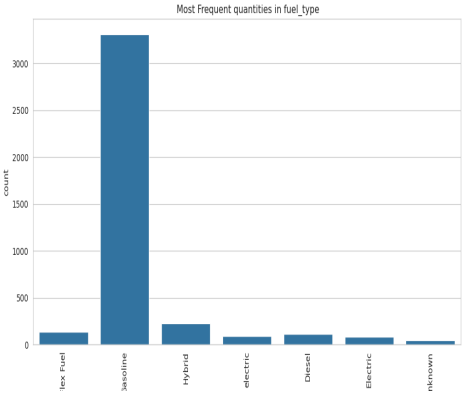


5. Frequency of some other columns:

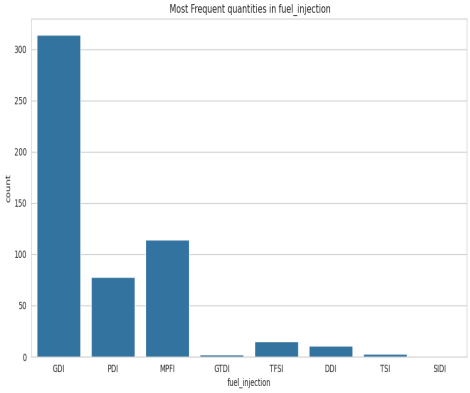


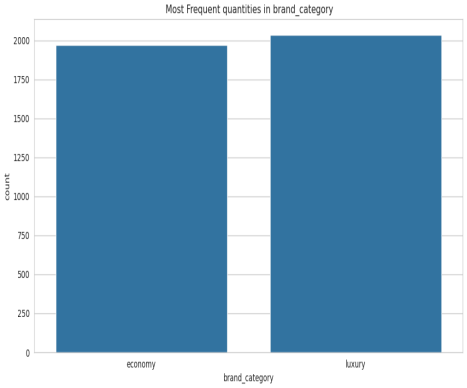




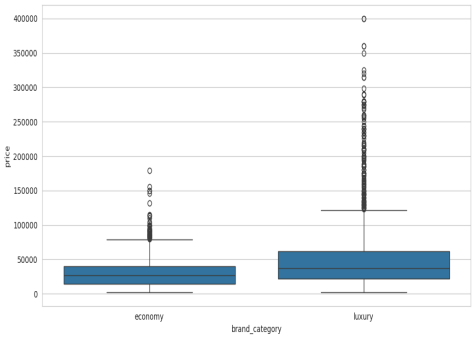


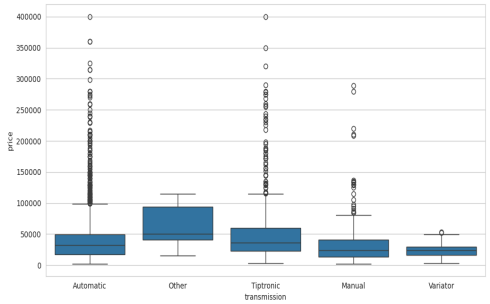
6. Both the luxury and economy brands have equal distribution.





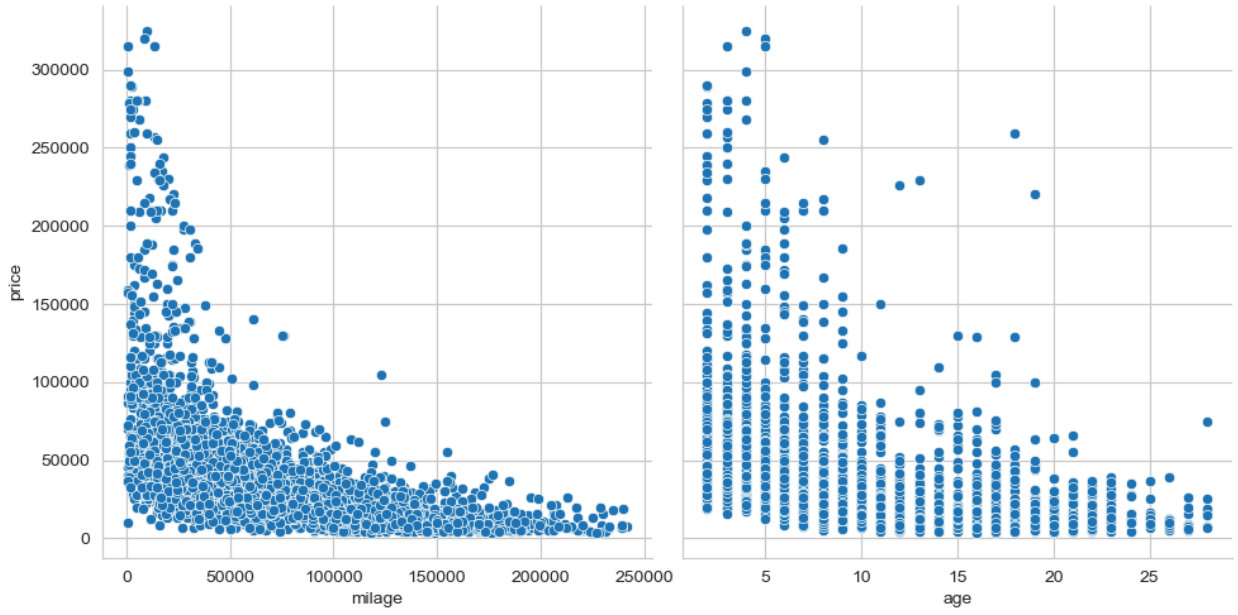
7. **Contribution to price:**





Song Gao:

1. **Dropping outliers**: I drop the data points that are out of 0.5%-99.5% for the feature's age, mileage, and price. After dropping the outliers, there are 3689 data points left.
2. The following scatter plots show how the price varies as mileage and age vary. We can see that as mileage and age increase, the price variable presents a trend of decreasing.



3. The following box plot shows how accident history affects the car price. It turns out that used cars without an accident history are more valuable.